Lecture 13: Variance, Chebyshev's Inequality, and Law of Large Numbers

CSCI2244-Randomness and Computation

April 1, 2019

The variance and standard deviation of a random variable measure how much the value of a random variable is likely to deviate from its mean–how 'spread out' it is.

1 Definition and three important properties.

If X is a random variable with $\mu = E(X)$, then

$$Var(X) = E((X-\mu)^2).$$

This is called the *variance* of X. We also define

$$\sigma(X) = \sqrt{Var(X)},$$

the *standard deviation* of X.

Just as $E(X)$ is not defined for every random variable X, $Var(X)$ might not be defined, even if $E(X)$ is defined.

When it is defined, we can use the linearity of expectation to derive:

$$\begin{aligned} Var(X) &= E((X-\mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

.

This is usually an easier way to compute the variance of X.

If c is a constant, then

$$\begin{aligned} Var(cX) &= E((cX)^2) - E(cX)^2 \\ &= E(c^2X^2) - (cE(X))^2 \\ &= c^2E(X^2) - c^2E(X)^2 \\ &= c^2(E(X^2) - E(X)^2 \\ &= c^2Var(X). \end{aligned}$$

Finally, suppose X, Y are independent random variables. As we've seen, this implies $E(XY) = E(X)E(Y)$. The following derivation repeatedly uses the linearity of expectation, and applies independence in the very last step.

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2 + 2XY + Y^2) - (E(X) + E(Y))^2 \\ &= (E(X^2) + 2E(XY) + E(Y^2)) - (E(X)^2 + 2E(X)E(Y) + E(Y^2)) \\ &= (E(X^2) - E(X)^2) + (E(Y^2) - E(Y)^2) + 2(E(XY) - E(X)E(Y)) \\ &= Var(X) + Var(Y) - 2 \times 0 \\ &= Var(X) + Var(Y) \end{aligned}$$

Again, don't forget that the hypothesis of independence is crucial for this additivity of variance to hold. (If it were not, then we would have $Var(2X) = Var(X + X) = 2 \cdot Var(X)$ for every random variable X, but we know $Var(2X) = 4 \cdot Var(X)$, which would give the odd result that every random variable has zero variance!)

2 Examples of computation of variance

2.1 Bernoulli random variable with parameter p

Let X be a Bernoulli random variable with parameter p. Since the values of X are just 0 and 1, $X = X^2$. Thus $E(X^2) = E(X) = p$, so

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p),$$

and thus

$$\sigma(X) = \sqrt{p(1-p)}.$$

Just for a reality check, let's recompute this using the original definition: $\mu = E(X) = p$, so $X - \mu$ has value $1 - p$ with probability p, and $-p$ with probability $1 - p$. So $(X - \mu)^2$ has value $(1 - p)^2$ with probability p, and p^2 with probability $(1 - p)$. Thus

$$E((X - \mu)^2) = p(1 - p)^2 + (1 - p)p^2 = (1 - p)(p(1 - p) + p^2) = (1 - p)p.$$

We got the same answer, of course. Observe that the method we used first is much easier.

2.2 Binomial random variable with parameters p, n.

The value of such a random variable X is the number of heads on n tosses of a coin with heads probability p. As we observed earlier, when talking about expected value,

$$X = X_1 + \cdots + X_n,$$

where each X_i is a Bernoulli random variable with parameter p, and the X_i are pairwise independent. Thus we can apply the additivity of variance, along with the previous result, and find

$$Var(X) = Var(X_1) + \cdots + Var(X_n) = np(1 - p),$$

and

$$\sigma(X) = \sqrt{np(1 - p)}.$$

If we take the *average* number of heads $Y = \frac{1}{n} \cdot X$, then we have

$$Var(Y) = \frac{1}{n^2} \cdot Var(X) = \frac{p(1 - p)}{n},$$

and

$$\sigma(Y) = \frac{\sqrt{p(1 - p)}}{\sqrt{n}}.$$

2.3 Spinner

For a continuous random variable X, we use the following fact to compute the variance:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 P_X(x) dx.$$

If X is the outcome of a single spinner, then $P_X(x)$ has the constant value 1 for x between 0 and 1, and is zero elsewhere, so

$$E(X^2) = \int_0^1 x^2 dx = \frac{1}{3}.$$

On the other hand, we already found that $E(X) = \frac{1}{2}$, so

$$Var(X) = \frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}.$$

2.4 Dartboard

Consider the random variable X that gives the distance of a dart from the center of a one-foot circular dartboard. We assume the darts are uniformly distributed. We found that the probability density function is

$$P_X(x) = \begin{cases} 0, & x < 0 \\ 2x, & 0 \leq x \leq 1 \\ 0, & x > 1. \end{cases}$$

Thus

$$E(X^2) = \int_0^1 2x^3 dx = \frac{1}{2}.$$

We already found $E(X) = \frac{2}{3}$, so

$$Var(X) = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}.$$

If you look at the graphs of the densities of the spinner and the dartboard example, you can kind of see that for the spinner, the probability mass is more ‘spread out’ on the interval $[0, 1]$ than for the dartboard, where it is more clustered toward 1, so it makes sense that the spinner gives a larger variance.

3 Markov’s inequality, Chebyshev’s inequality, and the Law of Large Numbers

The proportion of people in a population earning more than 3 times the average income cannot be greater than $\frac{1}{3}$. That seems sort of obvious. Let’s restate this principle in terms of probabilities: Let X be a random variable that takes on only nonnegative values, and let $\mu = E(X)$, and let > 0. Then

$$P(X \geq t\mu) \leq \frac{1}{t}.$$

In the example above, X is the salary of an individual selected uniformly at random from the population, and t, of course, is 3. This principle is called *Markov's inequality.* In class I gave a proof for discrete random variables; here is a proof in the continuous case.

$$\begin{aligned}
t\mu \cdot P(X \geq t\mu) &= t\mu \cdot \int_{t\mu}^{\infty} P_X(x)dx \\
&= \int_{t\mu}^{\infty} t\mu P_X(x)dx \\
&\leq \int_{t\mu}^{\infty} xP_X(x)dx \\
&\leq \int_{t\mu}^{\infty} xP_X(x)dx + \int_{0}^{t\mu} xP(x)dx \\
&= \int_{0}^{\infty} xP_X(x)dx \\
&= \mu,
\end{aligned}$$

which gives the result when we divide both sides of the inequality by $t\mu$. The step from the second to the third line is from the fact that

$$\int_{a}^{\infty} f(x)dx \leq \int_{a}^{\infty} g(x)dx$$

if $f(x) \leq g(x)$ for all $x \geq a$. We also used the fact that X takes on only positive values, so that μ itself is positive, and thus we can justify division of the inequality by $t\mu$.

Markov's inequality as it stands is not terribly useful, but it gives an important consequence when you apply it not to a random variable X itself, but to $Y = (X - \mu)^2$, so that $E(Y) = Var(Y)$. Then we get

$$P(Y \geq t^2 \cdot Var(Y)) \leq \frac{1}{t^2}.$$

The left-hand side is

$$P((X - \mu)^2 \geq t^2 \cdot Var(Y)) = P(|X - \mu| > t \cdot \sigma(Y)),$$

so for any random variable X for which the variance is defined, and any $t > 0$.

$$P(|X - \mu| > t \cdot \sigma(Y)) \leq \frac{1}{t^2}.$$

This is *Chebyshev's inequality.* It tells us, for example, that the probability that a random variable differs by more than 3 standard deviations from its mean is no more than $\frac{1}{9}$.

Example. Let's roll a single die and let X be the outcome. Then $E(X) = 3.5$. To compute $Var(X)$, we note

$$E(X^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6},$$

and thus

$$Var(X) = \frac{91}{6} - 3.5^2 \approx 2.92,$$

and

$$\sigma(X) \approx 1.71.$$

This tells us, for example, that

$$\begin{aligned} P(3 \leq X \leq 4) &= P(|X - \mu| \leq \frac{1}{2}) \\ &= P(|X - \mu| \leq \frac{\sigma(X)}{3.42}) \\ &= 1 - P(|X - \mu| > \frac{\sigma(X)}{3.42}) \\ &\geq 1 - 3.42^2 \\ &\approx -10.7. \end{aligned}$$

So Chebyshev's inequality told us that some probability is greater than -10.7, which means it told us nothing at all, since every probability is greater than or equal to 0.

But let's roll that die 100 times and repeat the calculation with the sum Y of the outcomes. Now $E(Y) = 350$, and $Var(Y) = 292$, $\sigma(Y) = \sqrt{292} \approx 17.1$. Now we have $50 \approx 2.92 \cdot \sigma(Y)$, so

$$\begin{aligned} P(300 \leq Y \leq 400) &= P(|Y - \mu| \leq 50) \\ &= P(|Y - \mu| \leq 2.92 \cdot \sigma(Y)) \\ &= 1 - P(|Y - \mu| > 2.92 \cdot \sigma(Y)) \\ &\geq 1 - 1/2.92^2 \\ &\approx 0.88. \end{aligned}$$

So there's at least an 88% probability that Y will be between 300 and 400. Now the inequality is really telling us something.

You can see where this is going: If we roll the die n times and let X be the sum, then the standard deviation is about $1.71\sqrt{n}$, and so

$$\begin{aligned} P(3n \leq X \leq 4n) &= P(|X-\mu| \leq n/2) \\ &= P(|X-\mu| \leq \frac{\sqrt{n}}{3.42} \cdot \sigma(X)) \\ &= 1 - P(|X-\mu| > \frac{\sqrt{n}}{3.42} \cdot \sigma(X)) \\ &\geq 1 - \frac{3.42^2}{n}. \end{aligned}$$

This obviously approaches 1 as a limit as the number of tosses gets larger. It's also obvious that there is nothing special about $3n$ and $4n$; any pair of bounds symmetrically spaced about the mean $3.5n$ would give the same result in the limit.

Here is the general principle: Suppose we have a random variable X for which the mean $\mu = E(X)$ and variance $\sigma(X)^2$ are defined. Let $\epsilon > 0$ be any positive number (think small). Now let let $Y_n = \frac{1}{n}(X_1 + \cdots + X_n)$, where the X_i are pairwise independent random variables having the same distribution as X—in other words, Y is the average of the results of n independent trials of whatever experiment led to X. Then $E(Y) = \mu$ and $\sigma(Y) = \frac{1}{n}\sigma(X)$. From Chebyshev's inequality,

$$\begin{aligned} P(|Y_n-\mu| > \epsilon) &= P(|Y_n-\mu| > \frac{\epsilon}{\sigma(Y)} \cdot \sigma(Y) \\ &\leq \frac{1}{(\epsilon/\sigma(Y))^2} \\ &= \frac{\sigma(Y)^2}{\epsilon^2} \\ &= \frac{\sigma(X)^2}{n^2\epsilon^2}. \end{aligned}$$

Even if ϵ is very small, so that $\frac{1}{\epsilon^2}$ is really, really big, if we make n large enough, the n^2 in the denominator will cause the right-hand side to approach 0 as n approaches ∞. Thus

$$\lim_{n\to\infty} P(|Y_n-\mu| > \epsilon) = 0.$$

In terms of complementary probability,

$$\lim_{n\to\infty} (\mu - \epsilon \leq Y_n \leq \mu + \epsilon) = 1.$$

This is called the *weak law of large numbers*. It tells us that the value Y_n approaches μ 'in probability': However small a deviation ϵ from the mean μ you name, if you perform the experiment often enough, the probability that its average value differs by as much as ϵ from the mean is vanishingly small.